



INVESTOR IN PEOPLE

The Patent Office
Concept House
Cardiff Road
Newport
South Wales
NP10 8QQ

I, the undersigned, being an officer duly authorised in accordance with Section 74(1) and (4) of the Deregulation and Contracting Out Act 1994, to sign and issue certificates on behalf of the Comptroller-General, hereby certify that annexed hereto is a true copy of the documents as originally filed in connection with the patent application identified therein together with the Statement of inventorship and of right to grant of a Patent (Form 7/77), which was subsequently filed.

In accordance with the Patents (Companies Re-registration) Rules 1982, if a company named in this certificate and any accompanying documents has re-registered under the Companies Act 1980 with the same name as that with which it was registered immediately before re-registration save for the substitution as, or inclusion as, the last part of the name of the words "public limited company" or their equivalents in Welsh, references to the name of the company in this certificate and any accompanying documents shall be treated as references to the name with which it is so re-registered.

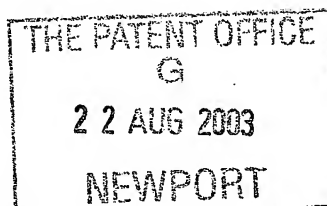
In accordance with the rules, the words "public limited company" may be replaced by p.l.c., plc, P.L.C. or PLC.

Re-registration under the Companies Act does not constitute a new legal entity but merely subjects the company to certain additional company law rules.

Signed

Dated 26 August 2003

**Statement of inventorship and of
right to grant of a patent**




7211603 BL0005-1 002564
The Patent Office

Cardiff Road
Newport
Gwent NP9 1RH

1. Your Reference
P.6952.GBA
2. Patent application number
(if you know it) 0219870.3
3. Full name of the or each applicant 20/20 Speech Limited
4. Title of the invention
Speech Synthesis Apparatus And Method
5. State how the applicant(s) derived the right
from the inventor(s) to be granted a patent
The applicant derived the right to be granted a patent
by virtue of the inventor's contract of employment
6. How many, if any, additional Patents Forms
7/77 are attached to this form?
(see note (c)) None
7. I/We believe that the person(s) named over the page (and on any extra
copies of this form) is/are the inventor(s) of the invention which the
above patent application relates to.

Signature Date: 19.08.2003


MAGUIRE BOSS
8. Name and daytime telephone number of
person to contact in the United Kingdom
PAUL J. EVENS Tel: 01480 301588

Notes

- a) If you need help to fill in this form or you have any questions, please contact the Patent Office on 0645 500505.
- b) Write your answers in capital letters using black ink or you may type them.
- c) If there are more than three inventors, please write the names and addresses of the other inventors on the back of another Patents Form 7/77 and attach it to this form.
- d) When an application does not declare any priority, or declares priority from an earlier UK application, you must provide enough copies of this form so that the Patent Office can send one to each inventor who is not an applicant.
- e) Once you have filled in the form you must remember to sign and date it.

Enter the full names, addresses and postcodes of the inventors in the boxes and underline the surnames

Roger Kenneth MOORE
20 Ebrington Road
West Malvern
Worc WR14 4NL
GB

Patents ADP number (if you know it)

08473498001

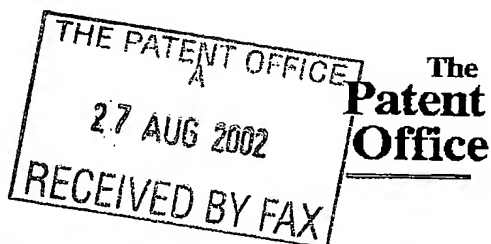
Patents ADP number (if you know it)

Patents ADP number (if you know it)

Reminder

Have you signed the form?

Patents Form 1/77

Patents Act 1977
(Rule 16)The
Patent
Office27AUG02 E743844-2 002824
P01/7700 0.00-0219870.3

Request for grant of a patent

(See the notes on the back of this form. You can also get an explanatory leaflet from the Patent Office to help you fill in this form).

The Patent Office
Cardiff Road
Newport
SOUTH WALES NP10 8QQ

1. Your Reference

P.6952.GBA

27 AUG 2002

2. Patent application number

(The Patent Office will fill in this part)

0219870.3

3. Full name, address and postcode of the or of each applicant (underline all surnames)

20/20 SPEECH LIMITED
Ixworth House
37 Ixworth Place
London SW3 3QH
G.B.

Patents ADP number (if you know it)

If the applicant is a corporate body, give the country/state of its incorporation

G.B.

07884961003

4. Title of the invention

SPEECH SYNTHESIS APPARATUS AND METHOD

5. Name of your agent (if you have one)

"Address for service" in the United Kingdom to which all correspondence should be sent (including the postcode)

MAGUIRE BOSS
5 Crown Street
St. Ives
Cambridgeshire
PE27 5EB, G.B.

Patents ADP number (if you know it)

07188725001

6. If you are declaring priority from one or more earlier patent applications, give the country and the date of filing of the or of each of these earlier applications and (if you know it) the or each application number

Country

Priority application number
(if you know it)Date of filing
(day/month/year)

7. If this application is divided or otherwise derived from an earlier UK application, give the number and the filing date of the earlier application

Number of earlier application

Date of filing
(day/month/year)

8. Is a statement of inventorship and of right to grant of a patent required in support of this request? (Answer 'Yes' if:

Yes

- a) any applicant named in part 3 is not an inventor, or
b) there is an inventor who is not named as an applicant, or
c) any named applicant is a corporate body;
See note (d)

Patents Form 1/77

Patents Form 1/77

9. Enter the number of sheets for any of the following items you are filing with this form.
Do not count copies of the same document

Continuation sheets of this form

Description	14 ✓
Claims(s)	3 ✓ <i>h</i>
Abstract	
Drawing(s)	3 <i>only</i> ✓

10. If you are also filing any of the following, state how many against each item.

Priority documents

Translations of priority documents

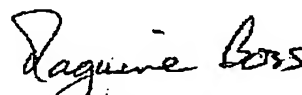
Statement of inventorship and right to grant of a patent (*Patents Form 7/77*)Request for preliminary examination and search (*Patents Form 9/77*)Request for substantive examination (*Patents Form 10/77*)Any other documents
(please specify)

11.

I/We request the grant of a patent on the basis of this application.

Signature

Date: 27/08/02



MAGUIRE BOSS

12. Name and daytime telephone number of person to contact in the United Kingdom

P.J. EVENS

Tel: 01480 301588

Warning

After an application for a patent has been filed, the Comptroller of the Patent Office will consider whether publication or communication of the invention should be prohibited or restricted under Section 22 of the Patents Act 1977. You will be informed if it is necessary to prohibit or restrict your invention in this way. Furthermore, if you live in the United Kingdom, Section 23 of the Patents Act 1977 stops you from applying for a patent abroad without first getting written permission from the Patent Office unless an application has been filed at least 6 weeks beforehand in the United Kingdom for a patent for the same invention and either no direction prohibiting publication or communication has been given, or any such direction has been revoked.

Notes

- a) If you need help to fill in this form or you have any questions, please contact the Patent Office on 0645 500505.
- b) Write your answers in capital letters using black ink or you may type them.
- c) If there is not enough space for all the relevant details on any part of this form, please continue on a separate sheet of paper and write "see continuation sheet" in the relevant part(s). Any continuation sheet should be attached to this form.
- d) If you have answered 'Yes' Patents Form 7/77 will need to be filed.
- e) Once you have filled in the form you must remember to sign and date it.
- f) For details of the fee and ways to pay please contact the Patent Office.

DUPLICATE

1

Speech synthesis apparatus and method

This invention relates to a speech synthesis apparatus and method.

- 5 The basic principle of speech synthesis is that incoming text is converted into spoken acoustic output by the application of various stages of linguistic and phonetic analysis. The quality of the resulting speech is dependent on the exact implementation details of each stage of processing, and the controls that are provided to the application programmer for controlling the synthesiser.
- 10 The final stage in a typical text-to-speech engine converts a detailed phonetic description into acoustic output. This stage is the main area where different known speech synthesis systems employ significantly different approaches. The majority of contemporary text-to-speech synthesis systems have abandoned traditional techniques based on explicit models of a typical human vocal tract in favour of concatenating
- 15 waveform fragments selected from studio recordings of an actual human talker. Context-dependent variation is captured by creating a large inventory of such fragments from a sizeable corpus of carefully recorded and annotated speech material. Such systems will be described in this specification as "concatenative".
- The advantage of the concatenative approach is that, since it uses actual recordings, it is
- 20 possible to create very natural-sounding output, particularly for short utterances with few joins. However, the need to compile a large database of voice segments restricts the flexibility of such systems. Vendors typically charge a considerable amount to configure a system for a new customer-defined voice talent, and the process to create such a bespoke system can take several months. In addition, by necessity, such systems
- 25 require a large memory resource (typically, 64-512 Mbytes per voice) in order to store as many fragments of speech as possible, and require significant processing power (typically 300-1000 MIPS) to perform the required search and concatenation.

For these reasons, concatenative TTS systems typically have a limited inventory of voices and voice characteristics. It is also the case that the intelligibility of the output of a concatenative system can suffer when a relatively large number of segments must be joined to form an utterance, or when a required segment is not available in the database.

- 5 Nevertheless, due to the natural sound of their output speech, such synthesisers are beginning to find application where significant computing power is available.

- A minority of contemporary text-to-speech synthesis systems continue to use a traditional formant-based approach that uses an explicit computational model of the resonances – formants – of the human vocal tract. The output signal is described by
10 several periodically generated parameters, each of which typically represents one formant, and an audio generation stage is provided to generate an audio output signal from the changing parameters. (These systems will be described as “parametric”.) This scheme avoids the use of recorded speech data by using manually derived rules to drive the speech generation process. A consequent advantage of this approach is that it
15 provides a very small footprint solution (1-5 Mbytes) with moderate processor requirements (30-50 MIPS). These systems are therefore used when limited computing power rules out the use of a concatenative system. However, the downside is that the naturalness of the output speech is usually rather poor in comparison with the concatenative approach, and formant synthesisers are often described as having a
20 ‘robotic’ voice quality, although this need not adversely affect the intelligibility of the synthesised speech.

An aim of this invention is to provide a speech synthesis system that provides the natural sound of a concatenative system and the flexibility of a formant system.

- From a first aspect, this invention provides a speech synthesiser having an output stage
25 for converting a phonetic description to an acoustic output, the output stage including a database of recorded utterance segments, in which the output stage:

- a. converts the phonetic description to a plurality of time-varying parameters;
- b. interprets the parameters as an key for accessing the database to identify an utterance segment in the database, and

c. outputs the identified utterance segment.

Thus, the parameters that are typically used to cause an output waveform to be generated and output instead cause a pre-recorded waveform to be selected and output. The parameters describe just a short segment of speech, so each segment stored in the database is small, so the database itself is small when compared with the database of a concatenative system. However, the database contains actual recorded utterances, which, the inventors have found, retain their natural sound when reproduced in a system embodying the invention.

Preferably, the output stage further comprises an output waveform synthesiser that can generate an output signal from the parameters. Such an output waveform synthesiser may be essentially the same as the parallel formant synthesiser used in a conventional parametric synthesis system. In the event that the parameters describe an utterance segment for which there is no corresponding recording in the database, the parameters can be passed to the output waveform synthesiser to generate an output signal. Thus, the synthesiser will operate in a concatenative mode where possible, and fall back to a parametric mode, as required.

In a synthesiser according to the last-preceding paragraph, the database can be populated to achieve an optimal compromise between memory requirements and perceived output quality. In the case of a synthesiser that is intended to generate arbitrary output, the larger the database, the greater the likelihood of operation in the concatenative mode. In the case of a synthesiser that is intended to be used predominantly or entirely to generate a restricted output repertoire, the database may be populated with segments that are most likely to be required to generate the output. For example, the database may be populated with utterance segments derived from speech by a particular individual speaker, by speakers of a particular gender, accent, and so forth. Of course, this restricts the range of output that will be generated in concatenative mode, but offers a reduction in the size of the database. However, it does not restrict the total output range of the synthesiser, which can always operate in parametric mode when required. It will be seen that selection of an appropriate database allows the implementation of an essentially continuous range of synthesisers

that achieve a compromise between quality and memory requirement most appropriate to a specific application.

In order that the database can be accessed quickly, it is advantageously an indexed database. In that case, the index values for accessing the database may be the values of
5 the time-varying parameters. Thus, the same values can be used to generate an output whether the synthesiser is operating in a concatenative mode or in a parametric mode.

The segments within the database may be coded, for example using linear predictive coding, GSM coding or other coding schemes. Such coding offers a system implementer further opportunity to achieve a compromise between the size of the
10 database and the quality of the output.

In a typical synthesiser embodying the invention, the parameters are generated in regular periodic frames, for example, with a period of several ms – more specifically, in the range 2 to 30 ms. For example, a period of approximately 10 ms may be suitable. In typical embodiments, there are ten parameters. The parameters may correspond to or
15 be related to speech formants. At each frame, an output waveform is generated, either from a recoding obtained from the database or by synthesis, these being reproduced in succession to create an impression of a continuous output.

From a second aspect, this invention provides a method of synthesising speech comprising:

- 20 a. generating from a phonetic description a plurality of time-varying parameters that describe an output waveform;
- b. interpreting the parameters to identify an utterance segment within a database of such segments that corresponds to the audio output defined by the parameters and retrieving the segment to create an output waveform; and
- 25 c. outputting the output waveform.

In a method embodying this aspect of the invention, if no utterance segment is identified in the database in step b, as corresponding to the parameters, an output waveform for output in step c is generated by synthesis.

Steps *a* to *c* are repeated in quick succession to create an impression of a continuous output. Typically, the parameters are generated in discrete frames, and steps *a* to *c* are performed once for each frame. The frames may be generated with a regular periodicity, for example, with a period of several ms – such as in the range 2 to 30 ms (e.g. 10 ms or thereabouts). The parameters within the frames typically correspond to or relate to speech formants.

In order to improve the perceived quality of output speech, it may be desirable not only to identify instantaneous values for the parameters, but also to take into account trends in the change of the parameters. For example, if several of the parameters are rising in value over several periods, it may not be appropriate to select an utterance segment that originated from a section of speech in which these parameter values were falling. Therefore, the output segment for any one frame may be selected as a function of the parameters of several frames. For example, the parameters of several surrounding frames may be analysed in order to create a set of indices for the database. While this may improve output quality, it is likely to increase the size of the database because there may be more than one utterance segment corresponding to any one set of parameter values. Once again, this can be used by an implementer as a further compromise between output quality and database size.

An embodiment of the invention will now be described in detail, by way of example, and with reference to the accompanying drawings, in which:

Figure 1 is a functional block diagram of a text-to-speech system embodying the invention;

Figure 2 is a block diagram of components of a text-to-speech system embodying the invention; and

Figure 3 is a block diagram of a waveform generation stage of the system of Figure 2.

Embodiments of the invention will be described with reference to a parameter-driven text-to-speech (TTS) system. However, the invention might be embodied in other types

of system, for example, including speech synthesis systems that generate speech from concepts, with no source text.

The basic principle of operation of a TTS engine will be described with reference to Figure 1. The engine takes an input text and generates an audio output waveform that
5 can be reproduced to generate an audio output that can be comprehended by a human as speech that, effectively, is a reading of the input text. Note that these are typical steps. A particular implementation of a TTS engine may omit one or more of them, apply variations to them, and/or include additional steps.

10 The incoming text is converted into spoken acoustic output by the application of various stages of linguistic and phonetic analysis. The quality of the resulting speech is dependent on the exact implementation details of each stage of processing, and the controls that the TTS engine provides to an application programmer.

15 Practical TTS engines are interfaced to a calling application through a defined application programmers interface (API). A commercial TTS engine will often provide compliance with the Microsoft (r. t. m.) SAPI standard, as well as the engine's own native API (that may offer greater functionality). The API provides access to the relevant function calls to control operation of the engine.

As a first step in the synthesis process the input text may be marked up in various ways in order give the calling application more control over the synthesis process (Step 110).
20 At present, several different mark-up conventions are currently in use, including SABLE, SAPI, VoiceXML and JSML, and most are subject to approval by W3C. These languages have much in common, both in terms of their structure and of the type of information they encode. However, many of the mark-up languages are specified in draft form only, and are subject to change. Presently, the most widely accepted TTS
25 mark-up standards are defined by Microsoft's SAPI and VoiceXML, but the "Speech Application Language Tags" has been commenced to provide a non-proprietary and platform-independent alternative.

As an indication of the purpose of mark-up handling, the following list outlines typical mark-up elements that are concerned with aspects of the speech output:

- Document identifier: identifies the XML used to mark up a region of text;
- Text insertion, deletion and substitution: indicates if a section of text should be inserted or replaced by another section;
- 5 • Emphasis: alters parameters related to the perception of characteristics such as sentence stress, pitch accents, intensity and duration;
- Prosodic break: forces a prosodic break at a specified point in the utterance;
- Pitch: alters the fundamental frequency for the enclosed text;
- Rate: alters the durational characteristics for the enclosed text;
- Volume: alters the intensity for the enclosed text;
- 10 • Play audio: indicates that an audio file should be played at a given point in the stream;
- Bookmark: allows an engine to report back to the calling application when it reaches a specified location;
- 15 • Pronunciation: controls the way in which words corresponding to the enclosed tokens are pronounced;
- Normalisation: specifies what sort of text normalisation rules should be applied to the enclosed text;
- Language: identifies the natural language of the enclosed text
- Voice: specifies the voice ID to be used for the enclosed text;
- 20 • Paragraph: indicates that the enclosed text should be parsed as a single paragraph;
- Sentence: indicates that the enclosed text should be parsed as a single sentence;

- Part of speech: specifies that the enclosed token or tokens have a particular part of speech (POS);
- Silence: produces silence in the output audio stream.

- 5 The text normalisation (or pre-processing) stage (112) is responsible for handling the special characteristics of text that arise from different application domains, and for resolving the more general ambiguities that occur in interpreting text. For example, it is the text normalisation process that has to use the linguistic context of a sentence to decide whether '1234' should be spoken as "one two three four" or "one thousand two hundred and thirty four", or whether 'Dr.' should be pronounced as "doctor" or "drive".
- 10 Some implementations have a text pre-processor optimised for a specific application domain (such as e-mail reading), while others may offer a range of pre-processors covering several different domains. Clearly, a text normaliser that is not adequately matched to an application domain is likely to cause the TTS engine to provide inappropriate spoken output.
- 15 The prosodic assignment component of a TTS engine performs linguistic analysis of the incoming text in order to determine an appropriate intonational structure (the up and down movement of voice pitch) for the output speech, and the timing of different parts of a sentence (step 114). The effectiveness of this component contributes greatly to the quality and intelligibility of the output speech.
- 20 The actual pronunciation of each word in a text is determined by a process (step 116) known as 'letter-to-sound' (LTS) conversion. Typically, this involves looking each word up in a pronouncing dictionary containing the phonetic transcriptions of a large set of words (perhaps more than 100 000 words), and employing a method for estimating the pronunciation of words that might not be found in the dictionary. Often TTS
- 25 engines offer a facility to handle multiple dictionaries; this can be used by system developers to manage different application domains. The LTS process also defines the accent of the output speech.

In order to model the co-articulation between one sound and another, the phonetic pronunciation of a sentence is mapped into a more detailed sequence of context-

dependent allophonic units (Step 118). It is this process that can model the pronunciation habits of an individual speaker, and thereby provide some 'individuality' to the output speech.

As will be understood from the description above, the embodiment shares features with a large number of known TTS systems. The final stage (Step 120) in a TTS engine converts the detailed phonetic description into acoustic output, and is here that the embodiment differs from known systems. In embodiments of the invention, a control parameter stream is created from the phonetic description to drive a waveform generation stage that generates an audio output signal. There is a correspondence between the control parameters and vocal formants.

The waveform generation stage of this embodiment includes two separate subsystems, each of which is capable of generating an output waveform defined by the control parameters, as will be described in detail below. A first subsystem, referred to as the "concatenative mode subsystem", includes a database of utterance segments, each derived from recordings of one or more actual human speakers. The output waveform is generated by selecting and outputting one of these segments, the parameters being used to determine which segment is to be selected. A second subsystem, referred to as the "parameter mode subsystem" includes a parallel formant synthesiser, as is found in the output stage of a conventional parameter-driven synthesiser. In operation, for each parameter frame, the waveform generations stage first attempts to locate an utterance segment in the database that best matches (according to some threshold criterion) the parameter values. If this is found, it is output. If it is not found, the parameters are passed to the parameter mode subsystem which synthesises an output from the parameter values, as is normal for a parameter driven synthesiser.

The structure of the TTS system embodying the invention will now be described with reference to Figure 2. Such a system may be used in implementations of embodiments of the invention. Since this architecture will be familiar to workers in this technical field, it will be described only briefly.

Analysis and synthesis processes of TTS conversion involve a number of processing. In this embodiment, these different operations are performed within a modular architecture

in which several modules 204 are assigned to handle the various tasks. These modules are grouped logically into an input component 206, a linguistic text analyser 208 (that will typically include several modules), a voice characterisation parameter set-up stage 210 for setting up voice characteristic parameters, a prosody generator 212, and a
5 speech sound generation group 214 that includes several modules, these being a converter 216 from phonemes to context-dependent PEs, a combining stage 218 for combining PEs with prosody, a synthesis-by-rule module 220, a control parameter modifier stage 222, and an output stage 224. An output waveform is obtained from the output stage 224.

10 In general, when text is input to the system, each of the modules takes some input related to the text, which may need to be generated by other modules in the system, and generates some output, which can then be used by further modules, until the final synthetic speech waveform is generated.

All information within the system passes from one module to another via a separate
15 processing engine 200 through an interface 202; the modules 204 do not communicate directly with each other, but rather exchange data bi-directionally with the processing engine 200. The processing engine 200 controls the sequence of operations to be performed, stores all the information in a suitable data structure and deals with the interfaces required to the individual modules. A major advantage of this type of
20 architecture is the ease with which individual modules can be changed or new modules added. The only changes that are required are in the accessing of the modules 204 in the processing engine; the operation of the individual modules is not affected. In addition, data required by the system (such as a pronouncing dictionary 205EI to specify how words are to be pronounced) tends to be separated from the processing
25 operations that act on the data. This structure has the advantage that it is relatively straightforward to tailor a general system to a specific application or to a particular accent, to a new language, or to implement the various aspects of the present invention.

The parameter set-up stage 210, includes voice characteristic parameter tables that define the characteristics of one or more different output voices. These may be derived
30 from the voices of actual human speakers, or they may be essentially synthetic, having characteristics to suit a particular application. A particular output voice characteristic

can be produced in two distinct modes. First, the voice characteristic can be one of those defined by the parameter tables of the voice characteristic parameter set-up stage 210. Second, a voice characteristic can be derived as a combination of two or more of those defined in the voice characteristic parameter set-up stage. The control parameter modifier stage 222 serves further to modify the voice characteristic parameters, and thereby further modify the characteristics of the synthesised voice. This allows speaker-specific configuration of the synthesis system. These stages permit characterisation of the output of the synthesiser to produce various synthetic voices, particularly deriving for each synthetic voice an individual set of tables for use in generating an utterance according to requirements specified at the input. Typically, the voice characteristic parameter set-up stage 210 includes multiple sets of voice characteristic tables, each representative of the characteristics of an actual recorded voice or of a synthetic voice.

As discussed, voice characteristic parameter tables can be generated from an actual human speaker. The aim is to derive values for the voice characteristic parameters in a set of speaker characterisation tables which, when used to generate synthetic speech, produce as close a match as possible, to a representative database of speech from a particular talker. In a method for generating the voice characterisation parameters, the voice characteristic parameter tables are optimised to match natural speech data that has been analysed in terms of synthesizer control parameters. The optimisation can use a simple grid-based search, with a predetermined set of context-dependent allophone units. There are various known methods and systems that can generate such tables, and these will not be described further in this specification.

Each voice characteristic parameter table that corresponds to a particular voice comprises a set of numeric data.

The parallel-formant synthesizer as illustrated in Figure 2 has twelve basic control parameters. These parameters are as follows:

Designation	Description
F0	Fundamental frequency
FN	Nasal frequency
F1, F2, F3	The first three formant frequencies
ALF, AL1 .. AL4	Amplitude controls
	Degree of voicing
	Glottal pulse open/closed ratio

Table 1

These control parameters are created in a stream of frames with regular periodicity, typically at a frame interval of 10 ms or less. To simplify operation of the synthesiser, some control parameters may be restricted. For example, the nasal frequency FN may
5 be fixed at, say, 250 Hz and the glottal pulse open/closed ratio is fixed at 1:1. This means that only ten parameters need be specified for each time interval.

Each frame of parameters is converted to an output waveform by a waveform generation stage 224. As shown in Figure 3, the waveform generation stage has a processor 310 (which may be a virtual processor, being a process executing on a
10 microprocessor). At each frame, the processor receives a frame of control parameters on its input. The processor calculates a database key from the parameters and applies the key to query a database 312 of utterance segments.

The query can have two results. First, it may be successful. In this event, an utterance segment is returned to the processor 310 from the database 312. The utterance segment
15 is then output by the processor, after suitable processing, to form the output waveform for the present frame. This is the synthesiser operating in concatenative mode.

Second, the query may be unsuccessful. This indicates that there is no utterance segment that matches (exactly or within a predetermined degree of approximation) the index value that was calculated from the keys. The processor then passes the
20 parameters to a parallel formant synthesiser 314. The synthesiser 314 generates an output waveform as specified by the parameters, and this is returned to the processor to be processed and output as the output waveform for the present claim. This is the synthesiser operating in parametric mode. Alternatively, the query may first be reformulated in an attempt to make an approximate match with a segment. In such
25 cases, it may be that one or more of the parameters is weighted to ensure that it is matched closely, while other parameters may be matched less strictly.

To generate an output that is perceived as continuous, successive output waveforms are concatenated. Procedures for carrying out such concatenation are well known to those skilled in the technical field. One such technique that could be applied in embodiments
30 of this invention is known as "pitch-synchronous overlap and add" (PSOLA). This is

fully described in Speech Synthesis and Recognition, John Holmes and Wendy Holmes, 2nd edition, pp 74-80, §5.4 onward. However, the inventors have found that any such concatenation technique must be applied with care in order that the regular periodicity of the segments does not lead to the formation of unwanted noise in the output.

- 5 In order to populate the database, recorded human speech is segmented to generate waveform segments of duration equal to the periodicity of the parameter frames. At the same time, the recorded speech is analysed to calculate a parameter frame that corresponds to the utterance segment.

- The recordings are digitally sampled (e.g. 16-bit samples at 22k samples per second).
10 They are then analysed (initially automatically by a formant analyser and then by optional manual inspection/correction) to produce an accurate parametric description at e.g. a 10 msec frame-rate. Each frame is thus annotated with (and thus can be indexed by) a set of (e.g. ten) parameter values. A frame corresponds to a segment of waveform (e.g. one 10 msec frame = 220 samples). During operation of the synthesiser, the same
15 formant values are derived from frames of the parameter stream to serve as indices that can be used to retrieve utterance segments from the database efficiently.

- If it is required to further compress the database at the expense of some loss of quality, the speech segments may be coded. For example, known coding systems such as linear predictive coding, GSM, and so forth may be used. In such embodiments, the coded
20 speech segments would need to be concatenated using methods appropriate to coded segments.

- In a modification to the above embodiment, a set of frames can be analysed in the process of selection of a segment from the database. The database lookup can be done using a single frame, or by using a set of (e.g. 3, 5 etc.) frames. For instance, trends in
25 the change of value of the parameters of the various frames can be identified, with the most weight being given to the parameters of the central frame. As one example, there may be two utterance segments in the database that correspond to one set of parameter values, one of the utterance segments being selected if the trend shows that the value of F2 is increasing and the other being selected if the value of F2 is decreasing.

The advantage of using a wider window (more frames) is that the quality of resulting match for the central target frame is likely to be improved. A disadvantage is that it may increase the size of the database required to support a given overall voice quality. As with selection of the database content described above, this can be used to optimise

5 the system by offsetting database size against output quality.

Claims

- 5 1. A speech synthesiser having an output stage for converting a phonetic description to an acoustic output, the output stage including a database of recorded utterance segments, in which the output stage:
 - a. converts the phonetic description to a plurality of time-varying parameters;
 - 10 b. interprets the parameters as an key for accessing the database to identify an utterance segment in the database, and
 - c. outputs the identified utterance segment.
2. A speech synthesiser according to claim 1 in which the output stage further comprises an output waveform synthesiser that can generate an output signal from the parameters.
- 15 3. A speech synthesiser according to claim 2 in which the output waveform synthesiser is essentially the same as the synthesiser used in a conventional parametric synthesiser.
4. A speech synthesiser according to claim 2 or claim 3 in which, in the event that the parameters describe an utterance segment for which there is no
20 corresponding recording in the database, the parameters are passed to the output waveform synthesiser to generate an output signal.
5. A speech synthesiser according to any one of claims 2 to 4 in which the database is populated to achieve an optimal compromise between memory requirements and perceived output quality.

6. A speech synthesiser according to claim 5 in which the database is populated with segments that are most likely to be required to generate a range of output corresponding to the application of the synthesiser.
- 5 7. A speech synthesiser according to claim 5 in which the database is populated with utterance segments derived from speech by a particular individual speaker, by speakers of a particular gender, accent, and so forth.
8. A speech synthesiser according to any preceding claim in which the database is an indexed database.
- 10 9. A speech synthesiser according to claim 8 in which the index values for accessing the database are the values of the time-varying parameters.
- 10 10. A speech synthesiser according to any preceding claim in which the segments within the database are coded, for example using linear predictive coding, GSM coding or other coding schemes.
- 15 11. A speech synthesiser according to any preceding claim in which the parameters are generated in regular periodic frames, for example, with a period of 2 to 30 ms.
12. A speech synthesiser according to claim 11 in which the period is approximately 10 ms.
- 20 13. A speech synthesiser according to claim 11 or claim 12 in which at each frame, an output waveform is generated these being reproduced in succession to create an impression of a continuous output.
14. A speech synthesiser according to any preceding claim in which the parameters correspond to or be related to speech formants.
- 25 15. A method of synthesising speech comprising:
 - a. generating from a phonetic description a plurality of time-varying parameters that describe an output waveform;

17

b. interpreting the parameters to identify an utterance segment within a database of such segments that corresponds to the audio output defined by the parameters and retrieving the segment to create an output waveform; and

5

c. outputting the output waveform.

16. A method of synthesising speech according to claim 15 in which, if no utterance segment is identified in the database in step *b*, as corresponding to the parameters, an output waveform for output in step *c* is generated by synthesis.

10

17. A method of synthesising speech according to claim 15 or claim 16 in which steps *a* to *c* are repeated in quick succession to create an impression of a continuous output.

18. A method of synthesising speech according to any one of claims 15 to 17 in which the parameters are generated in discrete frames, and steps *a* to *c* are performed once for each frame.

15

19. A method of synthesising speech according to claim 18 in which the frames are generated with a regular periodicity, for example, with a period of several ms (e.g. 10ms or thereabouts).

20. A method of synthesising speech according to claim 18 or claim 19 in which the parameters within the frames correspond to or relate to speech formants.

20

21. A speech synthesiser substantially as herein described with reference to the accompanying drawings.

22. A method of synthesising speech according to any one of claims 15 to 22 in which the output segment for any one frame are selected as a function of the parameters of several frames.

25

23. A method of synthesising speech substantially as herein described with reference to the accompanying drawings.

1/3

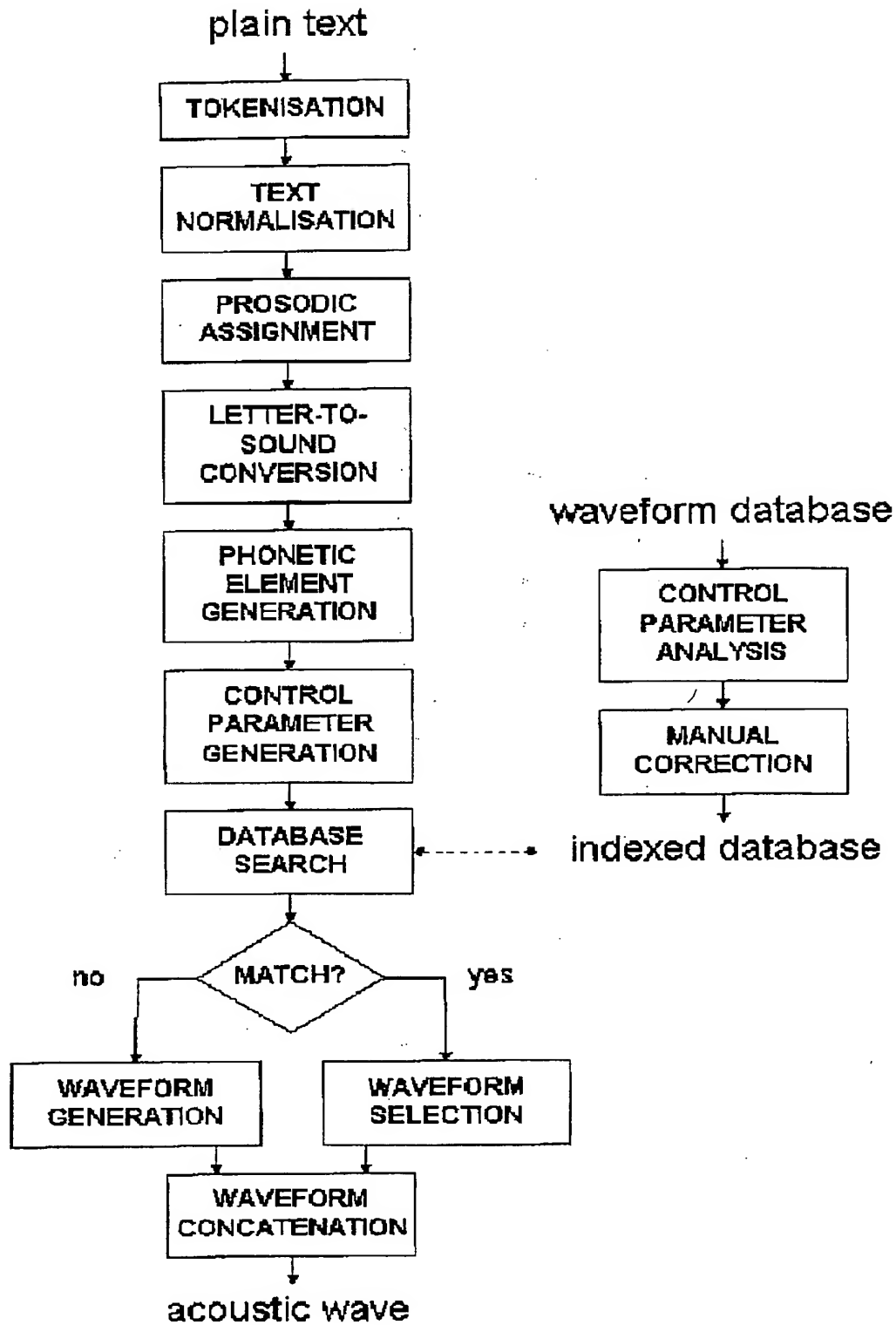


Fig. 1

2/3

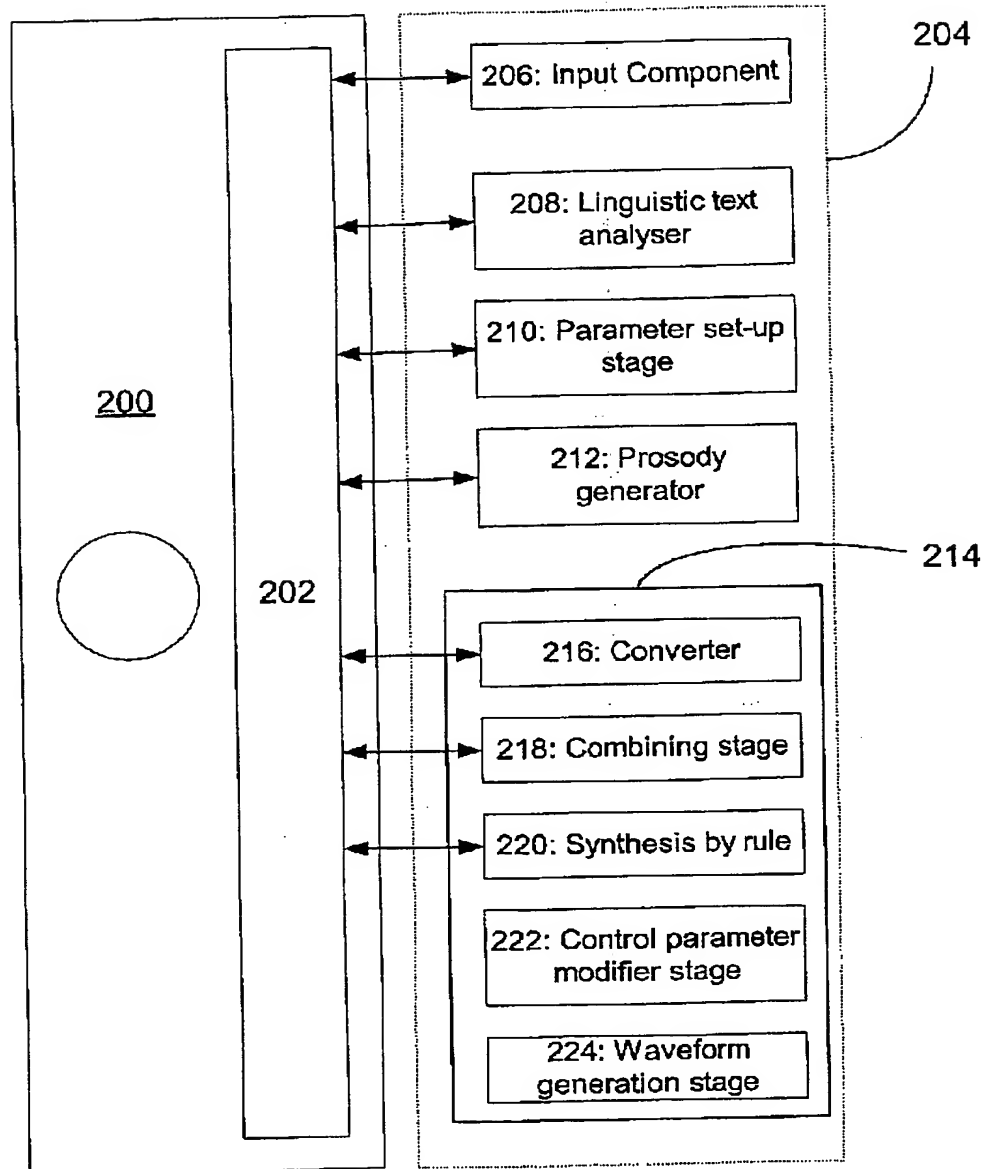


Fig 2

3/3

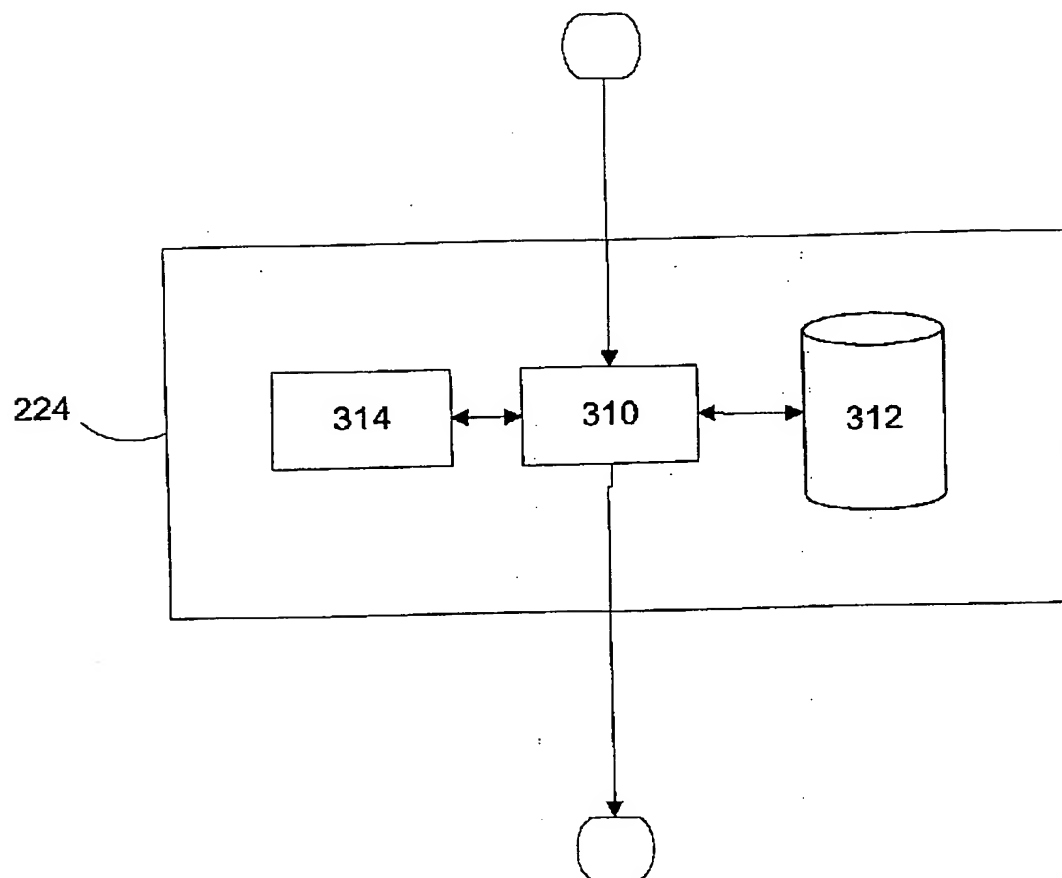


Fig 3